Regeneration in Markov Chain Samplers

By
Per Mykland[1], Luke Tierney[2] and Bin Yu[3]

Technical Report No. 585
School of Statistics
University of Minnesota
November 6, 1992

[1]Department of Statistics, University of Chicago. Research supported in part by grant DMS-8902667 and DMS-9204504 from the National Science Foundation.

[2]School of Statistics, University of Minnesota. Research supported in part by grant DMS-9005858 from the National Science Foundation.

[3]Department of Statistics, University of Wisconsin-Madison. Research supported in part by grant DAAL03-91-G-007 from the Army Research Office.

Abstract

Markov chain sampling has received considerable attention in the recent literature, in particular in the context of Bayesian computation and maximum likelihood estimation. This paper discusses the use of Markov chain splitting, originally developed as a tool for the theoretical analysis of general state space Markov chains, to introduce regeneration times into Markov chain samplers. This allows the use of regenerative methods for analyzing the output of these samplers, and can also provide a useful diagnostic of the performance of the samplers. The general approach is applied to several different samplers and is illustrated in a number of examples.

1 Introduction

In Markov chain Monte Carlo, a distribution π is examined by obtaining sample paths from a Markov chain constructed to have equilibrium distribution π. This approach was introduced by Metropolis et al. (1953) and has recently received considerable attention as a method for examining posterior distributions in Bayesian inference and for approximating the relative likelihood function in maximum likelihood estimation (see, for example, Tanner and Wong 1987, Gelfand and Smith 1990, Besag and Green 1993, Gilks et al. 1993, Smith and Roberts 1993, Tierney 1991a, Tierney 1991b, Liu et al. 1991, Geyer 1992, Geyer and Thompson 1992, and Yu 1992).

The analysis of the output produced by Markov chain samplers is more challenging than for other Monte Carlo methods, such as importance sampling, that are based on independent observations. The dependence in the samples makes estimating standard errors of Monte Carlo estimates more difficult. Furthermore, since it is usually not possible to start a Markov chain sampler with its equilibrium distribution, it may take some time for it to reach equilibrium, and it may therefore be useful to discard some initial portion of the sample to reduce the effect of the initial distribution used.

One approach to these problems is to try to identify *regeneration times* at which the chain restarts itself. The *tours* of the chain between regenerations are then independent. If the chain is observed for a fixed number of tours, then initialization issues are eliminated, and standard errors of sample path averages can be computed using methods based on *i.i.d.* observations. This approach is known as *regenerative simulation* (see, e.g., Ripley 1987, Section 6.4).

Regeneration times are easy to find for discrete Markov chains: if we fix a particular state, then the chain starts over every time it returns to that state. In general state space Markov chains, where the transition densities and the stationary distribution may be continuous, the chain may never return to any particular state. Nevertheless, several authors have developed ways of introducing regeneration times into general state space Markov chains (Athreya and Ney 1978, Nummelin 1978). The method of Nummelin (1978) is called *splitting*, and is well suited for use in regenerative simulation.

The paper is organized as follows. Section 2 reviews the regenerative simulation method. Section 3 first introduces the general splitting technique of Nummelin, and it then discusses ways to obtain better splits and proposes using regenerative simulation analysis as a diagnostic tool for Markov Chain samplers. Section 4 discusses the application of splitting to some Metropolis chains and Gibbs samplers, Section 5 illustrates these approaches using several examples, and Section 6 presents some final comments. Proofs of several results are given in the final section.

2 Regenerative Simulation

A stochastic process $\{X_n : n = 0, 1, \ldots\}$ is regenerative if there are times $T_0 \leq T_1 \leq \ldots$ such that at each T_i the future of the process is independent of the past. Then the tours of the process between these times are independent, and the times themselves form a renewal process. The renewal process is delayed if $T_0 \neq 0$.

Suppose the process has equilibrium distribution π, that we wish to estimate $\theta = E_\pi[f]$ for some function f, and that we observe the process for a fixed number n of tours. Let

$$N_i = T_i - T_{i-1}$$

and

$$Y_i = \sum_{j=T_{i-1}+1}^{T_i} f(X_j)$$

for $i = 1, \ldots, n$. Then the pairs (N_i, Y_i) are *i.i.d.*, and if $E[|Y_i|] < \infty$ and $E[N_i] < \infty$, then

$$\hat{\theta}_n = \frac{\sum Y_i}{\sum N_i} = \frac{\overline{Y}}{\overline{N}} \to \theta \tag{2.1}$$

by the law of large numbers and the renewal theorem. If the Y_i and N_i have finite variances, then the distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ converges to a $N(0, \sigma^2)$ distribution, and σ can be estimated using the variance estimation formula for a ratio estimator,

$$\hat{\sigma} = \frac{\sqrt{\sum (Y_i - \hat{\theta}_n N_i)^2}}{\sqrt{n}\,\overline{N}}. \tag{2.2}$$

If the number of tours is not large, then a jackknife estimate of variance may be more reliable (Ripley 1987, Section 6.4).

The motivation for regenerative simulation extends beyond asymptotics. Since

$$E\left[\sum Y_i\right] = \theta E\left[\sum N_i\right] \tag{2.3}$$

and

$$E\left[\sum (Y_i - \theta N_i)^2\right] = \mathrm{Var}\left(\sum Y_i - \theta N_i\right), \tag{2.4}$$

the bias in $\hat{\theta}_n$ is only due to ratio estimation, and the same is approximately true for $\hat{\sigma}^2$. Provided regenerations actually occur, there is no need to worry about the bias due to a slow mixing rate. This is true both for the estimate itself and for the estimate of uncertainty. Of course, means of *i.i.d.* samples can be problematic due to a heavy tailed distribution.

With regenerative simulation, the process is usually started with a regeneration, so $T_0 = 0$. If this is not possible, the simulation may have to be run from an arbitrary starting point until a regeneration occurs at some random time $T_0 \geq 0$. In this case the initial observations $X_0, \ldots, X_{T_0-1}$ are usually discarded.

A possible drawback of regenerative simulation is that using a fixed number of tours n leads to a random total run length T_n. This run length could be quite long if the time between regenerations is large and very variable. A regenerative simulation analysis can be used to estimate the variance of a simulation estimator even if the total run length is fixed at, say, t observations. If n_t is the number of complete tours within the first t observations, then an average over all t observations is close to the average over only the n_t complete tours. As a result, the standard error formula (2.2) remains asymptotically valid. Because of the waiting time paradox, the incomplete tour including the final observation is rather unusual. This leads to some biases that may be significant if the number of tours is small. Ripley (1987, Section 6.4) and Bratley, Fox and Schrage (1987, Sections 3.3.2 and 3.7) discuss these issues and give further references.

An intermediate strategy is to fix a target number t of observations, and to run the simulation until the next regeneration after t, i.e. to run $n_t + 1$ complete tours. This removes the waiting time paradox, and (2.3) and (2.4) remain valid.

3 Splitting Markov Chains

Let $\{X_n : n = 0, 1, \ldots\}$ be an irreducible Markov chain on a state space $(E, \mathcal{E})$ with transition kernel $P = P(x, dy)$ and invariant distribution π. The sigma algebra $\mathcal{E}$ is assumed to be countably generated. These assumptions imply that X_n is positive recurrent (see, for example, Tierney 1991b, Theorem 1). Assume in addition that X_n is Harris recurrent; this is satisfied by most Markov chain samplers (Tierney 1991b, Corollaries 1 and 2).

3.1 Nummelin's Splitting Technique

Suppose it is possible to find a function $s(x)$ and a probability measure $\nu(dy)$ such that

$$\pi(s) = \int s(x)\pi(dx) > 0$$

and

$$P(x, A) \geq s(x)\nu(A) \tag{3.1}$$

for all $x \in E$ and all $A \in \mathcal{E}$. Then the density

$$r(x, y) = \frac{s(x)\nu(dy)}{P(x, dy)} \tag{3.2}$$

exists and can be taken to satisfy $0 \leq r(x, y) \leq 1$. A pair (s, ν) satisfying these conditions is called a *split* for the transition kernel P.

Suppose the Markov chain X_n is generated as usual by first generating X_0, and then for each $n = 1, 2, \ldots$, generating X_{n+1} from $P(X_n, dy)$. To split this chain, a Bernoulli random variable S_n with success probability

$$P\{S_n = 1 | X_0, \ldots, X_{n+1}; S_0, \ldots, S_{n-1}\} = r(X_n, X_{n+1})$$

is generated for each $n = 0, 1, \ldots$. Thus S_n can be generated once X_{n+1} is available. Define

$$T_0 = \inf\{n \geq 0 : S_n = 1\}$$

and for each $n = 1, 2, \ldots$, let

$$T_n = \inf\{n > T_{n-1} : S_n = 1\},$$

with the convention that the infimum of the empty set is infinity. Then the T_n are regeneration times, and at each regeneration the Markov chain restarts with initial distribution ν:

Theorem 1 *Let X_n and S_n be constructed as described above. Then (X_n, S_{n-1}) is a Markov chain with transition kernel given by*

$$\begin{aligned} P\{X_{n+1} \in A, S_n = 1 | X_0, \ldots, X_n, S_0, \ldots, S_{n-1}\} &= s(X_n)\nu(A) \\ P\{X_{n+1} \in A, S_n = 0 | X_0, \ldots, X_n, S_0, \ldots, S_{n-1}\} &= P(X_n, A) - s(X_n)\nu(A) \end{aligned}$$

for any $A \in \mathcal{E}$. Hence $P\{S_n = 1 | X_n, S_{n-1}\} = s(X_n)$ and $P\{X_{n+1} \in A | X_n, S_{n-1}, S_n = 1\} = \nu(A)$. The regeneration times T_i are almost surely all finite, the equilibrium regeneration rate is

$$\lim_{n\to\infty} \frac{S_0 + \cdots + S_n}{n+1} = \pi(s) = \int s(x)\pi(dx) = \int r(x, y)\pi(dx)P(x, dy),$$

and the mean time between regenerations is

$$E[N_i] = E[T_i - T_{i-1}] = \frac{1}{\pi(s)}$$

for $i = 1, 2, \ldots$.

The construction outlined above only depends on a split (s, ν) through the product $s(x)\nu(dy)$. Thus it is not necessary to determine the normalizing constant needed to make ν into a probability measure. It is sufficient to find a finite, nonzero measure ν' and a function s' such that $s'(x)\nu'(dy) \leq P(x, dy)$ and $\pi(s') > 0$; then the split (s, ν) is given by $\nu(dy) = \nu'(dy)/\nu'(E)$ and $s(x) = s'(x)\nu'(E)$.

To apply the result of this theorem effectively, we need to be able to find a good split of a transition kernel P. Some useful approaches for specific types of samplers are discussed in Section 4: the remainder of this section examines general issues.

Splits of a given transition kernel P need not exist. If $\mathcal{E}$ is countably generated, Nummelin (1984) shows that it is always possible to find a split for the m-step transition kernel P^m for some $m \geq 1$. However, in Markov chain simulations P^m is rarely available in closed form for any $m > 1$, so we consider only the case $m = 1$.

If a split (s, ν) of P does exist, it is not unique: For any s' with $0 \leq s'(x) \leq s(x)$ for all x and $\pi(s') > 0$, the pair (s', ν) is also a split of P. But (s, ν) is a better choice, since it produces a smaller expected tour length. In general, given several choices for a split we usually prefer the ones that produce lower mean tour lengths, or higher regeneration rates. We will elaborate on this point in the following subsection.

If a split of a kernel is not available, then it may be possible to form a hybrid (Tierney 1991b, Section 2.4) with another kernel that is easier to split. If P_1 and P_2 are transition kernels with invariant distribution π, then the *cycle hybrid* kernel P_1P_2 and the *mixture hybrid* kernel $\alpha P_1 + (1-\alpha)P_2$ for $0 \leq \alpha \leq 1$ are also transition kernels with invariant distribution π. The cycle hybrid corresponds to alternately using P_1 and P_2 to generate a new state; for a mixture, at each step kernel P_1 is used with probability α and kernel P_2 with probability $1 - \alpha$. If a split of one of the kernels in a hybrid is available, then a split of the hybrid chain is available:

Proposition 1 *Suppose P_1 and P_2 are transition kernels with invariant distribution π and (s, ν) is a split for P_1. Then $(s, \nu P_2)$ is a split for the cycle kernel P_1P_2, and $(\alpha s, \nu)$ is a split for the mixture kernel $\alpha P_1 + (1-\alpha)P_2$ if $0 < \alpha \leq 1$.*

3.2 The Diagnostic Role of Regenerative Simulation

The regeneration rate $\pi(s)$ or its reciprocal, the mean tour length $E[N_i]$, and the pattern of the regeneration times can provide useful diagnostics for the performance of a Monte Carlo sampler. If $E[N_i]$ is small or the regeneration rate is high, then this suggests that the dependence between the observations produced by the sampler is rather mild. A more precise statement is given by the following proposition.

Proposition 2 *Let $\| \lambda \| = \max_{A \in \mathcal{E}} \lambda(A) - \min_{A \in \mathcal{E}} \lambda(A)$ denote the total variation norm of a signed measure λ, and suppose (s, ν) is a split for a transition kernel P. Then*

$$\| \pi(dx)P(x, dy) - \pi(dx)\pi(dy) \| \leq 2(1 - \{\pi(s)\}^2).$$

Thus the regeneration rate of a split provides a bound on the departure of the sampler from *i.i.d.* sampling.

A low regeneration rate by itself does not imply a highly dependent sequence. If a sampler is selecting *i.i.d.* observations from π, then (s, π) is a split for this *i.i.d.* Markov chain for any s with $0 \leq s(x) \leq 1$ for all x and $\pi(s) > 0$. Its tour lengths N_i are geometric random variables with expected value $\{\pi(s)\}^{-1}$. If (s, ν) is a split for a dependent sampler, then the mean tour length is also $E[N_i] = \{\pi(s)\}^{-1}$, but the distribution is not geometric; typically it will have a heavier upper tail.

For an *i.i.d.* sequence the probability of a regeneration occurring at any particular observation is a constant. As a result, the pattern of regenerations is uniform in the sense that the number of regenerations expected in a particular interval is proportional to the size of the interval. By the renewal theorem, a similar result is true asymptotically for any renewal process: If the process is observed for t periods, then for any a and b with $0 \leq a \leq b \leq 1$ the proportion of the observed regenerations that fall in the interval $[at, bt]$ converges to $(b-a)/E[N_i] = (b-a)\pi(s)$. On the other hand, a Markov chain sampler with a heavy-tailed regeneration distribution that is not observed for a sufficiently long time period may have occasional long periods with no regenerations that make the overall pattern of the renewals appear non-uniform.

These observations suggest that a plot of the regeneration pattern and an estimate of the density of regenerations per observation, a smoothed local regeneration rate, may be useful to assess the performance of a sampler. If a sampler is working well, then the regeneration pattern and the smoothed regeneration rate plot should be close to uniform. Departures from uniformity indicate that regenerations are less likely to occur when the sampler is in certain parts of the state space, and that it is taking considerable time, relative to the total observation time, to move from these regions into regions where regeneration is more likely to occur.

The regeneration rate can be estimated by the proportion of observations that result in regenerations. A smoothed regeneration rate plot can be produced as a histogram or a kernel density estimate of the observed regeneration times. Estimates of the regeneration rate and smoothed regeneration rate plots can also be constructed without directly using the split indicators by making use of the conditional regeneration rates; this can be viewed as a slight variance reduction by conditioning. The regeneration rate can be estimated using a sequence of length t by the average conditional regeneration rate,

$$\hat{r} = \frac{1}{t} \sum_{i=0}^{t-1} r(X_i, X_{i+1})$$

or, if the normalizing constant for ν is available, by

$$\tilde{r} = \frac{1}{t+1} \sum_{i=0}^{t} s(X_i).$$

In the case of Metropolis chains, one would typically use the estimate given by (4.3) below. The smoothed regeneration rate plot can be constructed by smoothing the $r(X_i, X_{i+1})$ or the $s(X_i)$.

If a low regeneration rate or a non-uniform regeneration pattern are observed, then it is worth exploring whether a better split can be obtained. Often this can be done as in Section 5 below by identifying a reasonable class of splits and then finding the best split in this family by maximizing the estimated regeneration rate or minimizing the mean tour length using preliminary samples or a normal approximation if one is available. If no improved split is found, the dependence structure in the sampler should be examined more closely to see if there are any suggestions for ways of reducing dependence, perhaps by constructing a hybrid sampler of the type described in Section 4.2.2

4 Splitting Some Markov Chain Samplers

Given a Markov chain sampler, three general approaches to incorporating regeneration are available. The first approach is to attempt to find a split for the sampler itself. This is possible for certain special samplers. If it is not possible to find a split for the original sampler, the second approach is to form a hybrid sampler that incorporates steps from a sampler for which a split is available. If the resulting chain does not regenerate very rapidly, then the third approach is to again choose a hybrid strategy, but one specifically designed to introduce more frequent regenerations. As discussed in Tierney (1991b), the Metropolis algorithm is particularly useful for constructing hybrid algorithms with particular properties.

4.1 Splitting Metropolis Chains

Suppose the distribution π we wish to sample has a density, also denoted by π, with respect to a measure μ, $\pi(dx) = \pi(x)\mu(dx)$. Hastings (1970) version of the Metropolis algorithm originally introduced by Metropolis et al. (1953) generates the next step X_{n+1} in a Markov chain from the current state X_n by first generating a candidate step Y from a transition kernel $Q(X_n, dy) = q(X_n, y)\mu(dy)$. This candidate is accepted with probability $\alpha(X_n, Y)$, where

$$\alpha(x, y) = \min\left\{\frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}, 1\right\}$$

and X_{n+1} is set equal to Y. Otherwise, the candidate is rejected and X_{n+1} is set equal to X_n.

For reasons to be discussed in connection with Proposition 3 below, it is natural to split a Metropolis kernel by splitting the sub-probability transition density $q(x, y)\alpha(x, y)$, i.e. by finding a pair (s', ν') such that

$$q(x, y)\alpha(x, y)\mu(dy) \geq s'(x)\nu'(dy). \tag{4.1}$$

Since the Metropolis kernel P satisfies

$$P(x, dy) \geq q(x, y)\alpha(x, y)\mu(dy),$$

this provides a split of the kernel P. The splitting variables S_n of Theorem 1 can be generated by allowing a split to occur only when a candidate step is accepted:

Theorem 2 *Suppose the Metropolis chain satisfies (4.1), and suppose (X_{n+1}, S_n) is generated as follows: (1) draw X_{n+1} conditionally on X_n by taking candidate steps from $q(X_n, y)\mu(dy)$ and accepting or rejecting according to $\alpha(X_n, y)$; (2) if the candidate in step (1) is rejected, then set $S_n = 0$; otherwise, generate S_n as a Bernoulli random variable with success probability $r_A(X_n, X_{n+1})$, where*

$$r_A(x, y) = \frac{s'(x)\nu'(dy)}{q(x, y)\alpha(x, y)}. \tag{4.2}$$

Then the process (X_{n+1}, S_n) has the distribution given in Theorem 1.

The success probability $r_A(x, y)$ is the conditional probability of a regeneration, given $X_n = x$ and $X_{n+1} = y$, and given that the candidate is accepted. In principle, it is possible to generate the splitting variables S_n directly using the success probability (3.2). But when ν' has atoms the expression for $r(x, x)$ can be complicated to derive due to the fact that $X_n = X_{n+1}$ can occur with

positive probability when the candidate step is accepted. When estimating the regeneration rate of a Metropolis chain, it is natural to use

$$\hat{r} = \frac{1}{t}\sum_{i=0}^{t-1} r_A(X_i, X_{i+1})\chi_{i+1} \tag{4.3}$$

where χ_{i+1} is the indicator function for the candidate step for X_{i+1} having been accepted.

This leaves the question of finding a split that satisfies (4.1). This can be done easily by finding a split (s_q, ν_q) for the kernel Q, provided there exists a positive function h such that

$$h(x)q(x,y) = h(y)q(y,x) \tag{4.4}$$

for all x and y. This condition is formally similar to a reversibility condition, but h is not required to be a probability density. This condition is satisfied by independence chains, where $q(x,y) = f(y)$ for some probability density f, by taking $h = f$. It is also satisfied by the original Metropolis algorithm, where it is assumed that $q(x,y) = q(y,x)$, by taking h to be a constant. If q satisfies (4.4) then the acceptance probability $\alpha(x,y)$ simplifies to

$$\alpha(x,y) = \min\left\{\frac{w(y)}{w(x)}, 1\right\}$$

where $w(x) = \pi(x)/h(x)$.

Theorem 3 *Suppose a Metropolis chain satisfies (4.4), and let (s_q, ν_q) be a split for Q. For any $c > 0$, set*

$$s'(x) = s_q(x)\min\left\{\frac{c}{w(x)}, 1\right\}$$

and

$$\nu'(dy) = \nu_q(dy)\min\left\{\frac{w(y)}{c}, 1\right\}.$$

Then (4.1) holds.

The ν' given in Theorem 3 is not a probability measure, but this does not matter in (4.1) as $\nu'(E)$ can be absorbed into s'.

The regeneration rate depends on the choice of the constant c. As mentioned in Section 3.2, a good choice for c can be determined by maximizing an estimate of the regeneration rate obtained either using an approximation to π, for example a normal approximation, or using a preliminary sample. It is not necessary to know the normalizing constant for π to determine a good choice of c; multiplying π by a constant is equivalent to dividing c by the same constant.

The reason for only permitting regenerations when the candidate step is accepted is twofold. On the one hand, this provides (through Theorems 2 and 3) a convenient formula for simulating the regeneration. On the other hand, very little is lost by having no regenerations at the times of rejection. This is because the dependence between X_n and X_{n+1} is so great in the case of rejection that only a very small probability of regeneration could be assigned anyway. This is particularly so if the Metropolis kernel is split with a ν that has no atoms, which must be the case if π has no atoms:

Proposition 3 *Let (s,ν) be a split for the Metropolis chain and assume that ν has no atoms. Then ν has a density with respect to μ, denoted by $\nu(y)$, such that*

$$q(x,y)\alpha(x,y) \geq s(x)\nu(y),$$

i. e. (s,ν) is a split for $q(x,y)\alpha(x,y)\mu(dy)$.

4.1.1 Independence Chains

In an independence Metropolis chain (Tierney 1991b, Section 2.3) candidates are generated from a fixed density f, regardless of the current state of the chain; thus $q(x,y) = f(y)$. Equation (4.4) holds for $h = f$. To apply Theorem 3, choose $s_q = 1$ and $\nu_q(dy) = f(y)\mu(dy)$. This independence chain will be used to form hybrid chains in Section 4.2.2.

For an independence chain with the split of Theorems 2 and 3, the distribution ν' has density proportional to

$$f(y)\min\left\{\frac{w(y)}{c},1\right\} = f(y)\min\left\{\frac{\pi(y)}{cf(y)},1\right\}.$$

This can be sampled by rejection sampling to obtain an initial value X_0 for the chain corresponding to a regeneration. The conditional probability (4.2) of a regeneration at step n, given $X_n = x$, $X_{n+1} = y$, and no rejection, simplifies to

$$\begin{aligned} r_A(x,y) &= \frac{\min\left\{\frac{w(y)}{c},1\right\}\min\left\{\frac{c}{w(x)},1\right\}}{\min\left\{\frac{w(y)}{w(x)},1\right\}} \\ &= \begin{cases} \dfrac{c}{\min\{w(x),w(y)\}} & \text{if } w(x) > c \text{ and } w(y) > c \\ \dfrac{\max\{w(x),w(y)\}}{c} & \text{if } w(x) < c \text{ and } w(y) < c \\ 1 & \text{otherwise} \end{cases} \end{aligned} \tag{4.5}$$

Thus a regeneration is certain to have occurred if $w(x)$ and $w(y)$ are on opposite sides of c. This suggests that a good choice for c will typically be in the center of the distribution of the weights $w(x)$ under π.

If the candidates for an independence chain are produced by a rejection algorithm with candidate generation density proportional to g (Tierney 1991b, Section 2.3) and if $c = 1$, then this regeneration probability becomes

$$r_A(x,y) = \begin{cases} 1 & \text{if } x \in C \text{ or } y \in C \\ \min\left\{\dfrac{g(x)}{\pi(x)},\dfrac{g(y)}{\pi(y)}\right\} & \text{otherwise,} \end{cases}$$

where $C = \{x : \pi(x) \le g(x)\}$ is the set where g dominates π. Thus a regeneration occurs whenever X_n is at a point where g is an envelope for π.

Simple expressions for the equilibrium regeneration rate of an independence chain are also available:

Proposition 4 *For an independence chain with candidate generation density f and the split given above, the equilibrium regeneration rate is*

$$\begin{aligned} \pi(s) &= c\left(\int \min\left\{\frac{f(x)}{\pi(x)},\frac{1}{c}\right\}\pi(x)\mu(dx)\right)^2 \\ &= c\left(\int \min\left\{1,\frac{1}{c}\frac{\pi(x)}{f(x)}\right\}f(x)\mu(dx)\right)^2 \\ &= c\left(\int_{\{x:w(x)<c\}}\frac{1}{c}\pi(x)\mu(dx) + \int_{\{x:w(x)\ge c\}} f(x)\mu(dx)\right)^2 \end{aligned} \tag{4.6}$$

These expression can be used to estimate the regeneration rate or to select a good choice of c based on either an approximation to π or a preliminary sample. To use this expression to estimate the regeneration rate, the normalizing constant for π must be available or must be estimated. For an un-normalized density $\pi(x)$, the first identity in Equation (4.6) becomes

$$\pi(s) = \frac{c}{\int \pi(x)\mu(dx)} \left(\int \min\left\{ \frac{f(x)}{\pi(x)}, \frac{1}{c} \right\} \pi(x)\mu(dx) \right)^2.$$

From this expression it is clear that it is not necessary to estimate the normalizing constant to select a minimizing value of c.

If π and f are $N(0, A)$ and $N(0, B)$ distributions, respectively, then the region of integration for $\pi(x)$ in the third identity in Equation (4.6) is

$$\{X : w(X) < c\} = \left\{ X : Y < \log\left(c \frac{|A|^{1/2}}{|B|^{1/2}} \right) \right\}$$

with

$$Y = \frac{1}{2} X^T \left(B^{-1} - A^{-1} \right) X.$$

The distribution of the quadratic form Y can be represented as a linear combination of squares of independent standard normals. This can be used for the approximate computation of c if π and f are approximately normal.

4.1.2 Standard Metropolis Chains

The original version of the Metropolis algorithm of Metropolis et al. (1953) assumes a symmetric candidate generation kernel, $q(x, y) = q(y, x)$. In this case, equation (4.4) holds with h a constant. To apply Theorem 3, a split (s_q, ν_q) of q has to be found.

One useful approach to finding such a split is to select as the distribution ν_q the candidate generation density for some distinguished point $\tilde{x}$, so $\nu_q(dy) = q(\tilde{x}, y)\mu(dy)$, and then determine $s_q(x)$ as

$$s_q(x) = \inf_{y \in E} \frac{q(x, y)}{q(\tilde{x}, y)}.$$

Unfortunately this sometimes produces a function that is identically zero and thus not acceptable. A minor modification is to choose a convenient set $D \in \mathcal{E}$, usually a compact set, and to take

$$\nu_q(dy) = \frac{q(\tilde{x}, y) 1_D(y) \mu(dy)}{\int_D q(\tilde{x}, u)\mu(du)}$$

and

$$s_q(x) = \inf_{y \in D} \frac{q(x, y)}{q(\tilde{x}, y)}.$$

It is possible to start the chain with a regeneration by rejection sampling the initial state X_0 from a density proportional to $q(\tilde{x}, y) 1_D(y)$.

As an example, consider a random walk chain with normal increments, so

$$q(x, y) \propto \exp\left\{ -\frac{1}{2}(y - x)^T (y - x) \right\},$$

and let $D = \{y : |y| \leq d\}$ for some $d > 0$. Then for $\tilde{x} = 0$

$$s_q(x) = \inf_{y \in D} \frac{q(x,y)}{q(\tilde{x},y)} = \exp\left\{-\frac{1}{2}x^T x - d|x|\right\}.$$

A similar approach can be used for any random walk chain based on a spherically symmetric increment distribution.

4.2 Regeneration and the Gibbs Sampler

Suppose the state space E is a product of d components, $E = E_1 \times \cdots \times E_d$, an element of E is written as $x = (x_1, \ldots, x_d)$ with $x_i \in E_i$, and $\pi(x)$ is a density with respect to a product measure $\mu(dx) = \mu(dx_1) \times \cdots \times \mu(dx_d)$. Let $\pi_i(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$ denote the conditional density of the i-th component given all the others. The Gibbs sampler (Gelfand and Smith 1990) starting with $X_n = x$ generates X_{n+1} by generating

$$\begin{aligned} X_{n+1,1} &= y_1 \quad \text{from} \quad \pi_1(y_1|x_2, \ldots, x_d) \\ X_{n+1,2} &= y_2 \quad \text{from} \quad \pi_2(y_2|y_1, x_3, \ldots, x_d) \\ &\vdots \\ X_{n+1,d} &= y_d \quad \text{from} \quad \pi_d(y_d|y_1, \ldots, y_{d-1}) \end{aligned}$$

Even though examples of very strong dependence are available, experience suggests that for many problems the dependence in the Gibbs sampler sequence drops off very quickly, often within 10 to 20 cycles. As a result, it seems reasonable that a regeneration scheme with mean tour lengths on the order of 10 to 20 should be available in these cases.

For some problems it is possible to identify a split of the Gibbs sampler itself. For others, it is easier to form a hybrid algorithm of the Gibbs sampler and independence Metropolis steps, and to use the approach of the preceding section to identify regenerations that occur on the independence chain steps.

4.2.1 Splitting a Gibbs Sampler

The transition kernel of the Gibbs sampler has transition density

$$p(x,y) = \pi_1(y_1|x_2, \ldots, x_d)\pi_2(y_2|y_1, x_3, \ldots, x_d) \cdots \pi_d(y_d|y_1, \ldots, y_{d-1}). \tag{4.7}$$

In some cases it may be possible to find a split of this density by direct examination. For others, it may be possible to follow the strategy used for splitting the standard Metropolis transition kernel by choosing a distinguished point $\tilde{x}$ and a set $D \in \mathcal{E}$, taking $\nu(dy)$ to have density $p(\tilde{x}, y)$, and setting

$$s(x) = \inf_{y \in D} \frac{p(x,y)}{p(\tilde{x},y)}. \tag{4.8}$$

In many problems the minimization required to compute $s(x)$ can take advantage of the exponential family structure often present in problems where a Gibbs sampler is used; this is the case for one of the examples discussed below in Section 5.

The computation of $s(x)$ may also be simplified by a suitable choice of the ordering of the components. For example, the final factor in (4.7) does not depend on x and therefore cancels from the ratio in the definition of $s(x)$. For the purposes of computing a split, it is therefore useful to place the most complicated conditional distribution last in the update sequence. In the case

of two-components, i.e. for $d = 2$, a split of π_1 or π_2 is thus automatically a split of the Gibbs sampler.

Once again, starting the chain with a regeneration is easy, since sampling from $\nu(dy)$ corresponds to taking one Gibbs sampler cycle starting at $\tilde{x}$.

4.2.2 Hybrid Samplers

Tierney (1991b) describes several strategies for forming hybrid samplers by combining several more basic samplers. In particular, it is often useful to combine Gibbs samplers with steps from an independence Metropolis chain. If a Gibbs sampler is producing highly correlated observations, then the use of independence steps with a well chosen candidate generation density can help to reduce these correlations. Whether the use of independence steps helps or not depends on the quality of the candidate generation density. Examining the regeneration pattern produced by a split of the independence steps can provide a useful indication of the success of these steps. On the other hand, if a Gibbs sampler does seem to be performing well, using periodic independence steps can help to confirm the performance of the sampler and to provide regeneration points for a regenerative analysis.

The problem of choosing a good density for the independence steps is similar to the problem of choosing a good importance sampling density or a good initial density for a Gibbs sampler. Split t distributions (Geweke 1989) and the over-dispersed distributions of Gelman and Rubin (1992) may be useful. We may also use the normal approximation to the posterior distribution, if one is available, or use the density corresponding to one Gibbs cycle from some reasonable initial point, $\tilde{x}$. If it is possible to sample from the prior distribution and the likelihood is bounded, then choosing an independence kernel that samples from the prior distribution with positive probability produces an independence kernel with bounded weight function w. The computations required for performing alternate Gibbs cycles and independence steps are roughly equivalent to the computations required for identifying a split using the approach outlined above. But there is no need to carry out the careful analysis needed to determine the split of a Gibbs sampler; we only have to do a limited amount of experimenting with preliminary samples or approximate posterior distributions to identify a good choice of the constant c in Theorem 3.

It is of course not necessary to alternate independence and Gibbs steps. Instead, a sampler could use an independence step after every five or ten Gibbs steps. This increases the lengths of the average tours, but should still result in a reasonable number of tours in most simulations, provided the Gibbs sampler is in fact working reasonably well. The independence steps can also be incorporated as a mixture rather than as a cycle.

When the independence step candidate distribution f is less spread out than the target distribution π, the independence steps do not contribute significantly to the mixing of the chain. Instead, they can be viewed as simply checking whether the chain has returned to the region containing the bulk of the mass of f.

5 Examples

5.1 A Hierarchical Poisson Model

One of the examples presented by Gelfand and Smith (1990) is a hierarchical Poisson model. Failures in ten pumps at a nuclear power plant are assumed to occur according to independent Poisson processes with each pump having its own failure rate $\lambda_1, \ldots, \lambda_{10}$. The pumps were observed for periods t_i of varying lengths, and the numbers of observed failures s_i for each pump were

recorded. The data were originally analyzed in Gaver and O'Muircheartaigh (1987) and are also reproduced in Tierney (1991b, Table 1). Conditional on a hyperparameter β, the individual pump failure rates are assumed to be independent random variables with a gamma distribution $G(\alpha, \beta)$ with density proportional to $x^{\alpha-1}e^{-\beta x}$. The hyperparameter β has a gamma distribution $G(\gamma, \delta)$ with $\gamma = 0.01$ and $\delta = 1$. For the gamma exponent of the rate distribution Gelfand and Smith use the method of moments estimator, $\alpha = 1.802$.

For the resulting posterior distribution, given β, the λ_i are independent $G(\alpha + s_i, t_i + \beta)$ random variables, and, given $\lambda_1, \ldots, \lambda_{10}$, the distribution of β is $G(\gamma + 10\alpha, \sum \lambda_i + \delta)$. For constructing a split of the Gibbs sampler, suppose we first generate β, then $\lambda_1, \ldots, \lambda_{10}$. Then new values of β and λ_i only depend on the previous ones through $\Lambda = \sum \lambda_i$, and the ratio of the Gibbs sampler transition densities started with two different values of Λ only depends on the next state through its value of β:

$$\frac{p(x,y)}{p(\tilde{x},y)} = \left(\frac{\Lambda(x)+\delta}{\Lambda(\tilde{x})+\delta}\right)^{\gamma+10\alpha} \exp\{(\Lambda(\tilde{x}) - \Lambda(x))\beta(y)\}.$$

In this equation x and y represent different combinations of β and λ's. To apply the approach outlined in Section 4.2.1, we need to choose a distinguished value $\tilde{x}$, or its corresponding value of Λ, $\tilde{\Lambda} = \Lambda(\tilde{x})$, and a set D, which only needs to depend on β, and compute

$$s(\Lambda) = P\{\beta \in D|\tilde{\Lambda}\} \inf_{\beta \in D} \left(\frac{\Lambda+\delta}{\tilde{\Lambda}+\delta}\right)^{\gamma+10\alpha} \exp\{(\tilde{\Lambda} - \Lambda)\beta\}$$

For an interval $D = [d_1, d_2]$, the minimization produces

$$s(\Lambda) = P\{d_1 \le \beta \le d_2|\tilde{\Lambda}\} \left(\frac{\Lambda+\delta}{\tilde{\Lambda}+\delta}\right)^{\gamma+10\alpha} \exp\{(\tilde{\Lambda} - \Lambda)d(\Lambda)\},$$

where

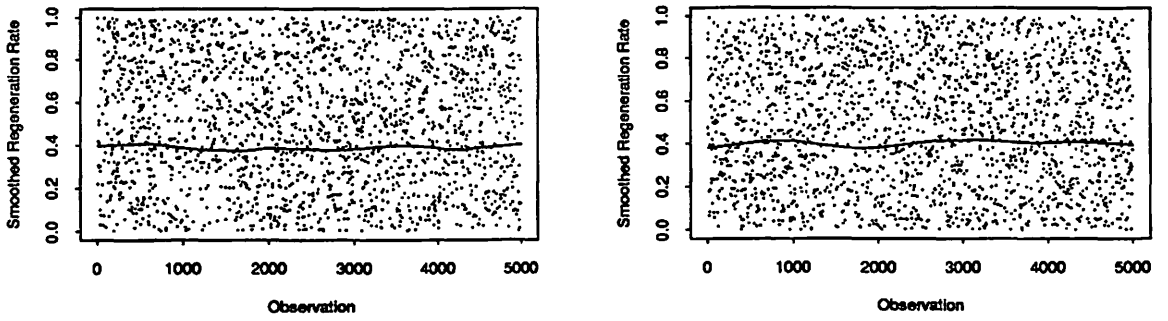
$$d(\Lambda) = \begin{cases} d_1 & \text{if } \Lambda < \tilde{\Lambda} \\ d_2 & \text{if } \Lambda \ge \tilde{\Lambda}. \end{cases}$$

The corresponding conditional probability of a regeneration, given $X_n = x$ and $X_{n+1} = y$, is

$$r(x,y) = \begin{cases} \exp\{(\tilde{\Lambda} - \Lambda(x))(d_1 - \beta(y))\} & \text{if } \Lambda(x) < \tilde{\Lambda} \text{ and } d_1 \le \beta \le d_2 \\ \exp\{(\tilde{\Lambda} - \Lambda(x))(d_2 - \beta(y))\} & \text{if } \Lambda(x) \ge \tilde{\Lambda} \text{ and } d_1 \le \beta \le d_2 \\ 0 & \text{otherwise.} \end{cases}$$

A reasonable approach to choosing the three parameters $\tilde{\Lambda}$, d_1 and d_2 of this split is to set $\tilde{\Lambda}$ equal 6.7, the approximate posterior mean of Λ based on a preliminary sample, and to choose d_i of the form $\tilde{\beta} \pm k\tilde{s}_\beta$, where $\tilde{\beta} = 2.35$ and $\tilde{s}_\beta = 0.69$ are the approximate posterior mean and standard deviation of β, again based on a preliminary sample. By graphing estimated regeneration rates, based on the preliminary sample, for a range of values of k, the optimal value of k was found to be 1.1; this corresponds to choosing $d_1 = 1.6$ and $d_2 = 3.1$. These choices are essentially identical to ones obtained by a global optimization over all three parameters. Using this split on a Gibbs sampler run of length 5000 produced an estimated regeneration rate of $\hat{r} = 0.39$, or an estimated expected tour length of $1/0.39 = 2.56$. The number of observed regenerations was 1968, corresponding to 1967 complete tours. Figure 1(a) shows a smoothed regeneration rate plot as well as a jitter plot of the observed regeneration times for this sampler run.

As a second approach, we used an alternating sampler in which a Gibbs cycle was followed by an independence step. The independence step candidates were generated by a single Gibbs cycle



(a) Gibbs Sampler (b) Alternating Gibbs/Independence Sampler

Figure 1: Smoothed regeneration rate plots for the pump data example for sampler runs of length 5000 from a pure Gibbs sampler and an alternating Gibbs/independence sampler. The plot backgrounds show jitter plots of the observed regeneration times.

started with $\tilde{\Lambda} = 6.7$, the approximate posterior mean of Λ, and generating first β and then the λ_i. The weight function for this independence kernel is

$$w(x) = \frac{e^{\tilde{\Lambda}\beta(x)}}{\prod(t_i + \beta(x))^{s_i+\alpha}}.$$

This can be standardized to be near one by dividing by its value at $\tilde{\beta} = 2.35$, the estimated posterior mean of β based on a preliminary sample. Using a preliminary Gibbs sample of 100 with Equation (4.6) and the standardized weights produced an estimated optimal value of $c = 1.1$ for the constant in Theorem 3; the estimated regeneration rate for the independence steps at this value of c was 0.78. The regeneration rate for a chain that alternates Gibbs steps and independence steps should thus be approximately $0.78/2 = 0.39$. A run of 5000 from this alternating chain, consisting of 2500 Gibbs cycles and 2500 independence steps, produced an estimated regeneration rate of $\hat{r} = 0.42$, or an estimated tour length of $1/0.42 = 2.38$. A total of 2070 regenerations, or 2069 complete tours, were observed. Figure 1(b) shows a smoothed regeneration rate plot for this Gibbs/independence sampler along with a jitter plot of the observed regeneration times.

Both the split Gibbs sampler and the alternating sampler have very high regeneration rates and uniform regeneration patterns that confirm that the Gibbs sampler works very well in this problem. Since the independence steps used here only use a single Gibbs step to generate their candidates, they do not significantly accelerate convergence of the algorithm; but additional acceleration does not seem necessary in this case. In computational cost, one full Gibbs/independence cycle is close to one Gibbs step with a split generation; it could therefore be argued that a split Gibbs sampler run of length 5000 should be compared to an alternating run of length 10000. On the other hand, neither the split generation nor the independence steps need to be carried out on every cycle; using them, say, every 5-th cycle would reduce the cost of checking for regenerations and increase the expected tour lengths from approximately 2.5 to approximately 12.5.

5.2 A Normal Variance Components Model

Gelfand et al. (1990, Section 4) consider a variance components model where

$$Y_{ij} = \theta_i + e_{ij}, \quad i = 1, \ldots, K, \quad j = 1, \ldots, J,$$

with e_{ij}, given the θ_i and σ_e^2, having independent $N(0, \sigma_e^2)$ distributions, and the θ_i, given μ and the variances, being independent $N(\mu, \sigma_\theta^2)$ random variables. The parameters μ, σ_θ^2 and σ_e^2 are independent $N(\mu_0, \sigma_0^2)$, $IG(a_1, b_1)$, and $IG(a_2, b_2)$ random variables, respectively. Here $IG(a, b)$ denotes an inverse gamma distribution.

Gelfand et al. use a Gibbs sampler based on the conditional distributions

$$\begin{aligned}
\sigma_\theta^2 | Y, \mu, \theta, \sigma_e^2 &= IG\left(a_1 + \frac{1}{2}K, b_1 + \frac{1}{2}\sum(\theta_i - \mu)^2\right) \\
\sigma_e^2 | Y, \mu, \theta, \sigma_\theta^2 &= IG\left(a_2 + \frac{1}{2}KJ, b_2 + \frac{1}{2}\sum\sum(Y_{ij} - \theta_i)^2\right) \\
\mu | Y, \theta, \sigma_\theta^2, \sigma_e^2 &= N\left(\frac{\sigma_\theta^2\mu_0 + \sigma_0^2\sum\theta_i}{\sigma_\theta^2 + K\sigma_e^2}, \frac{\sigma_\theta^2\sigma_e^2}{\sigma_\theta^2 + K\sigma_e^2}\right) \\
\theta_i | Y, \mu, \sigma_\theta^2, \sigma_e^2 &= N\left(\frac{J\sigma_\theta^2\overline{Y}_i + \sigma_e^2\mu}{J\sigma_\theta^2 + \sigma_e^2}, \frac{\sigma_\theta^2\sigma_e^2}{J\sigma_\theta^2 + \sigma_e^2}\right),
\end{aligned}$$

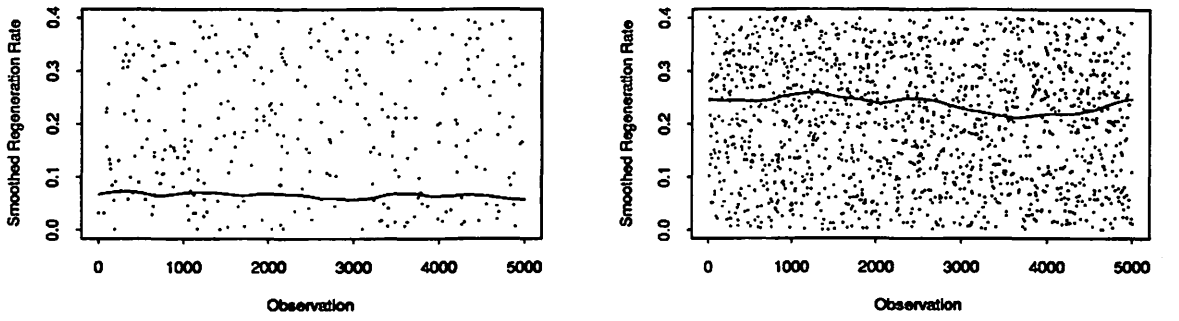
with the θ_i conditionally independent given Y, μ, σ_θ^2, and σ_e^2. It would also be possible to combine the θ_i and μ, since their conditional distribution, given the data and the other parameters, is a multivariate normal distribution. This should result in a more efficient Gibbs sampler, but for comparison we used the decomposition of Gelfand et al. (1990).

The data set used by Gelfand et al. is an artificial data set introduced by Box and Tiao (1973, p. 247). This data set was generated from the model with $\mu = 5$, $\sigma_\theta^2 = 4$ and $\sigma_e^2 = 16$. The prior distribution parameters used for μ and σ_e^2 were $\mu_0 = 0$, $\sigma_0 = 10^{12}$, and $a_2 = b_2 = 0$. For σ_θ^2 Gelfand et al. consider two cases, I: $a_1 = b_1 = 0$ and II: $a_1 = 1/2$ with several values of b_1; we consider only the case case $b_1 = 1$. Both these prior distributions are improper because of the choice of prior distribution for σ_e^2.

It is possible to use the ideas of Section 4.2.1 to split the Gibbs sampler for this problem. But the algebra is rather messy, so we will only apply the hybrid approach of Section 4.2.2. We use an alternating Gibbs/independence chain with the Gibbs step generating first new values for σ_θ^2 and σ_e^2, then μ, and finally new θ_i's. The independence candidate density is a single Gibbs step started at the average of 100 preliminary runs of the pure Gibbs sampler.

For prior I, the optimal choice of c based on a preliminary sample of 100 and Equation (4.6) was approximately $c = e^{-7}$, with an estimated regeneration rate of 0.020. But in a run of 5000 with the alternating sampler, a split using this value of c produced only two regenerations and an estimated regeneration rate of 0.0005. This in itself does not imply that there is a problem, but further examination shows that the posterior distribution for this prior is in fact improper (Hill 1965). Any irreducible Markov chain sampler for this posterior will be either null recurrent or dissipative, so the lack of frequent regeneration is not surprising.

For prior II, a preliminary Gibbs sample of 100 gave an estimated regeneration rate for independence steps of 0.152 at $c = e^{-2.7}$. Using a run of 5000 of the alternating sampler, the estimated regeneration rate was 0.066, corresponding to an estimated mean tour length of 15.15. A total of 332 regenerations, or 331 complete tours, were observed. Figure 2(a) shows the smoothed regeneration rate plot and a jitter plot of the regeneration times.



(a) Gibbs Step Candidate Dist. (b) Approx. Marginal Candidate Dist.

Figure 2: Smoothed regeneration rate plots for the variance components example with prior II for sampler runs of length 5000 from two alternating Gibbs/independence samplers. The plot backgrounds show jitter plots of the observed regeneration times.

These results suggest that the Gibbs sampler is working reasonably well. But it may be possible to improve the sampler by using a better candidate generation density in the independence steps. If we could sample μ, σ_θ^2, and σ_e^2 from their joint marginal posterior distribution, then we would only need to draw θ_i's from their conditional distribution given the other parameters to obtain an *i.i.d.* sample from the posterior distribution. This suggests that we may be able to improve our candidate generation density by matching the distribution used to generate μ, σ_θ^2, and σ_e^2 more closely to their marginal distribution. As a very simple attempt, we tried generating them independently with σ_θ^2 and σ_e^2 chosen from log normal distributions and μ from a t distribution with five degrees of freedom. The parameters for these distributions were estimated from the preliminary run of length 100 of the Gibbs sampler. Using Equation (4.6) and the preliminary sample, the optimal c was approximately $c = 1$ with an estimated regeneration rate for the independence steps of 0.632. Using a sample of 5000 from the alternating sampler, the estimated regeneration rate was 0.293, giving an estimated expected tour length of 3.41, and the number of observed regenerations was 1481. The smoothed regeneration rate plot and a jitter plot of the regeneration times are show in Figure 2(b).

This example illustrates several points. The regenerative simulation analysis can give an indication of problems in a sampler, and it can also be used to indicate when there is room for improvement in the sampler. For many hierarchical models a Gibbs sampler can be improved by combining it with independence steps that sample the hyperparameters from an approximation to their joint marginal posterior distribution. In this example a simple match of features of the one dimensional margins of the three parameters was sufficient to produce a significant improvement. In other cases it might be useful to use kernel density estimates of the marginal distribution, or other approaches that take advantage of features in the joint distribution.

5.3 An Artificial Example

As an artificial example that illustrates the diagnostic performance of regenerative simulation, we considered a distribution π that is a mixture of a bivariate standard normal distribution and a bivariate standard normal distribution shifted to have its center at the point (μ, μ). The mixing

μ	Tours	Regen. Rate	$\overline{X}_1$	$\overline{X}_2$	$\mathrm{SE}_B(\overline{X}_1)$	$\mathrm{SE}_R(\overline{X}_1)$	$\mathrm{SE}_B(\overline{X}_2)$	$\mathrm{SE}_R(\overline{X}_2)$
1	1458	0.289	0.516	0.493	0.018	0.019	0.020	0.020
3	1289	0.258	1.449	1.458	0.061	0.070	0.062	0.068
5	1312	0.262	2.374	2.355	0.218	0.375	0.219	0.381
7	989	0.198	4.233	4.214	0.342	1.671	0.346	1.691

Table 1: Summary statistics for Gibbs/independence samplers for mixtures of bivariate normal densities. Standard errors were computed using batch means (SE_B) and the regenerative method (SE_R).

probability was 0.5. For sufficiently large μ the density π is bimodal with the modes displaced along the diagonal; the Gibbs sampler should therefore have some difficulty in moving from one mode to the other. An alternating Gibbs/independence hybrid sampler was constructed with a single Gibbs step from the origin as the candidate generation density for the independence steps. This example is intended to model a situation where preliminary exploration has revealed one mode, the mode at the origin, but a second, equally important, mode is in fact present at (μ, μ).

Runs of length 5000 were performed for μ equal to 1, 3, 5, and 7. Table 1 shows the number of complete tours, the estimated equilibrium regeneration rate, and the sample means of the two coordinates. Two estimated standard errors are given for each sample mean, a batch mean estimate based on batches of size 50, and a regenerative estimate based on (2.2). The means of the marginal distributions of the two coordinates under π are equal to $\mu/2$. Figure 3 shows the smoothed regeneration rate plots as well as jitter plots of the observed regeneration times for the four values of μ. As expected, the performance of the sampler deteriorates as μ increases. At $\mu = 5$ there are several large gaps in the regeneration times, corresponding to periods when the sampler is in the mode at (μ, μ). For $\mu = 7$ the sampler starts in the mode at the origin, moves to the second mode after approximately 800 observations, and returns to the mode at the origin after a total of approximately 3800 observations. The estimated standard errors are also considerably larger for $\mu = 5$ and $\mu = 7$.

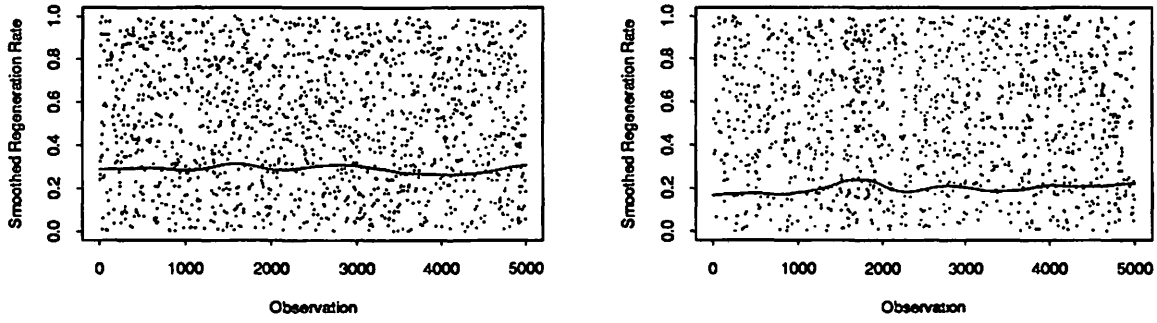
The regenerative simulation analysis clearly reveals that the sampler is not behaving well for $\mu = 5$ and $\mu = 7$. In a real example, further exploration should reveal the second mode. Incorporating this mode into a candidate generation density for independence steps in a hybrid sampler should produce a sampler with much better properties.

5.4 Splitting and the Swendsen–Wang Algorithm

Even though our work is primarily motivated by applications in Bayesian and maximum likelihood computations, the ideas can also be used in other Markov chain Monte Carlo problems. As an illustration, we show how they can be applied to the Swendsen–Wang algorithm.

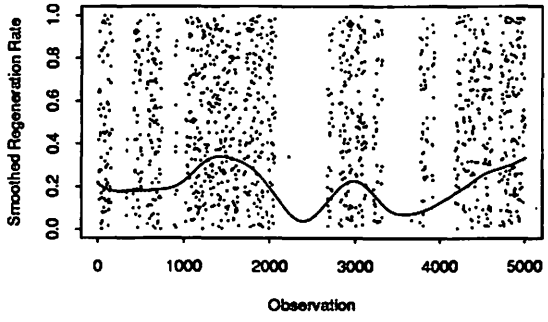
Swendsen and Wang (1987) propose a method for sampling the Potts (1952) model (Besag and Green 1993), the multicolor generalization of the Ising model. This model assumes the vertices $V = \{1, \ldots, M\}$ of a graph (V, E) are each given one of L colors, x_i. The distribution of the colors is assumed proportional to $\exp\{-\beta\eta(x)\}$, where $\eta(x)$ is the number of edges $(i, j) \in E$ for which $x_i \neq x_j$, and β is a non-negative constant.

Given the colors x_i, the algorithm adds auxiliary *bond variables*, b_{ij}. No bonds are placed between vertices with different colors. If $x_i = x_j$ and (i, j) is an edge in the graph, then with probability $1 - \exp(-\beta)$ a bond is placed between vertices i and j, and $b_{ij} = 1$. Otherwise, no

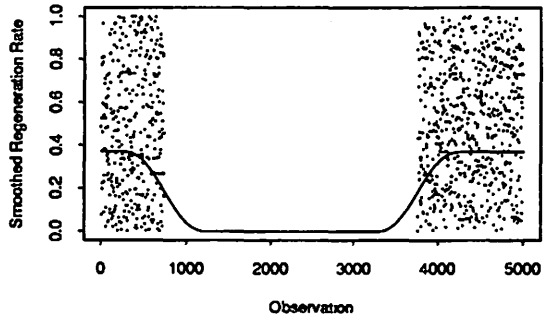


(a) $\mu = 1$

(b) $\mu = 3$



(c) $\mu = 5$



(d) $\mu = 7$

Figure 3: Smoothed regeneration rate plots for the bivariate normal mixture example based on runs of length 5000 from an alternating Gibbs/independence sampler. The Plot backgrounds show jitter plots of the observed regeneration times.

bond is placed between the vertices, and $b_{ij} = 0$. A set of bonds partitions the vertex set V into connected components. Conditional on the bonds b_{ij}, the x_i's are the same within components, and the component colors are selected independently and uniformly from the available L colors. The joint distribution of (x, b) is proportional to

$$e^{-\beta(|E|-\sum b_{ij})}(1-e^{-\beta})^{\sum b_{ij}}$$

on the set of (x, b) values such that $b_{ij} = 0$ whenever $x_i \neq x_j$, and is zero elsewhere; here $|E|$ is the number of edges in the graph, and sums are over edges. The algorithm is a two-coordinate Gibbs sampler that alternates between selecting bonds and colors from these conditional distributions.

It is possible to split this Gibbs sampler along the lines of Section 4.2.1. A natural choice for the distinguished state $\tilde{x}$ is the state where all vertices have the same color; the sampler then generates a new set of bonds as *i.i.d.* Bernoulli random variables, and then selects a new set of colors for the resulting components. In computing the infimum needed to find the splitting probability $s(\cdot)$ in Equation 4.8, taking $D = E$, it is easy to see that any difference in color produces an infimum of zero. So $s(\cdot)$ will be just the indicator of whether all vertices have the same color or not, and the split will occur each time the sampler returns to a configuration in which all states have the same color.

An alternating sampler using independence steps can also be constructed using Gibbs steps started at a uniform color configuration as the candidate generator. Since the conditional distribution of the colors given bonds is just uniform on the $L^{c(b)}$ possible component colorings, where $c(b)$ is the number of components, the weight function for this independence step is proportional to $L^{c(b)}$.

Both the split of the Gibbs sampler itself and the split of the alternating chain should work reasonably well if β is not too small and the graph not too large. As a simple illustration, we used an $m \times m$ grid with $m = 32$ and two colors, $L = 2$. The parameter β was chosen to make the bond placement probability $(1 - e^{-\beta})$ equal to 0.8. (This corresponds to a temperature well below the freezing point of the infinite Ising lattice; a more elaborate candidate generation density would be needed for lower values of β or higher values of m.) Both samplers were started with all vertices the same color.

For the pure Gibbs sampler with a split on returns to all one color, using a run of 5000 gave 197 regenerations, or 196 complete tours; this gives an estimated regeneration rate of 0.039 and an estimated mean tour length of 25.6. Using a preliminary Gibbs sampler run of length 100, the optimal choice of c for the independence split was found to be $c = e^4$, with an estimated regeneration rate for the independence steps of 0.166. Based on a run of 5000 of the alternating chain, the estimated rate was 0.073, corresponding to a mean tour length of 13.7; there were 358 regenerations, or 357 complete tours.

6 Conclusions

Identifying regeneration points in a Markov Chain sampler eliminates initialization issues and allows variance estimates to be computed based on *i.i.d.* observations. In addition, it allows Markov chain sampling to take advantage of a parallel computing environment without the problems created by many short Markov chain runs when regeneration points are not available.

Regeneration rates and regeneration time distributions are also useful as a diagnostic of sampler performance. High regeneration rates and regeneration patterns that are close to uniform suggest that the sampler is working well. Low rates or non-uniform regeneration patterns do not necessarily

imply that there is a problem, but do suggest that the dependence in the sampler is worth examining more closely. Of course, as with *i.i.d.* sampling in general, there is no guarantee based on only a finite run of a sampler that a longer run would not produce several very long tours that might correspond to explorations of additional modes of the posterior distribution. Constructing a good density for use in independence steps can help to reduce this problem, but it cannot eliminate it.

In principle, the methods outlined in this paper can be used to split samplers whenever the single step transition density is available; this is the case for most samplers proposed for exploring posterior distributions. The resulting splits will not always be satisfactory — there are many reasonable samplers for which the basic mixing rate is too slow to provide reasonable splits based on a single transition. A hybrid algorithm that incorporates independence steps may help to accelerate mixing; the regenerative analysis can be used to asses whether this attempt at acceleration has succeeded. Success depends on the quality of the candidate generation density. Good methods for choosing these densities are available for many problems, but more work is needed in developing methods for deriving satisfactory densities for high dimensional problems. Adaptive methods for producing these densities seem particularly worthy of further exploration.

7 Proofs

Proof of Theorem 1. The construction of (X_n, S_{n-1}) is given in Nummelin (1984, pages 61-62). Corollary 4.2 of Nummelin shows that the recurrence of X_n implies that the renewal sequence T_i is recurrent, i.e. that all regeneration times are finite. Convergence of the observed regeneration rate follows from Theorem 3 of Tierney (1991b), and the expression for the mean time between regenerations is given on page 76 of Nummelin. □

Proof of Proposition 1. For the cycle kernel,

$$\begin{aligned}(P_1P_2)(x,dy) &= \int P_1(x,du)P_2(u,dy)\\ &\geq s(x)\int \nu(du)P_2(u,dy)\\ &= s(x)(\nu P_2)(dy);\end{aligned}$$

thus $(s, \nu P_2)$ is a split of P_1P_2. For the mixture kernel,

$$\alpha P_1(x,dy) + (1-\alpha)P_2(x,dy) \geq \alpha P_1(x,dy) \geq \alpha s(x)\nu(dy).$$

So $(\alpha s, \nu)$ is a split of $\alpha P_1 + (1-\alpha)P_2$ as long as $\alpha > 0$. □

Proof of Proposition 2. Since π is invariant for P, integrating the split inequality (3.1) with respect to $\pi(dx)$ shows that $\pi(dy) \geq \pi(s)\nu(dy)$. As a result, ν is absolutely continuous with respect to π with a density g such that $0 \leq \pi(s)g(x) \leq 1$ for all x. Therefore

$$\begin{aligned}\| \pi(dx)P(x,dy) - \pi(dx)\pi(dy) \| &\leq \| \pi(dx)P(x,dy) - \pi(dx)s(x)\nu(dy) \|\\ &\quad + \| \pi(dx)s(x)\nu(dy) - \pi(dx)\pi(dy) \|\\ &= 1 - \pi(s) + \int\int |1 - s(x)g(y)|\pi(dx)\pi(dy).\end{aligned}$$

Now

$$\begin{aligned}\int\int |1 - s(x)g(y)|\pi(dx)\pi(dy) &\leq \int |1 - s(x)|\pi(dx) + \int\int s(x)|1 - g(y)|\pi(dx)\pi(dy)\\ &= 1 - \pi(s) + \pi(s)\int |1 - g(y)|\pi(dy)\end{aligned}$$

and

$$\begin{aligned}\int |1-g(y)|\pi(dy) &= \int |1-\pi(s)g(y)-(1-\pi(s))g(y)|\pi(dy) \\ &\leq \int |1-\pi(s)g(y)|\pi(dy)+1-\pi(s) \\ &= 2(1-\pi(s))\end{aligned}$$

Combining these results produces

$$\| \pi(dx)P(x,dy)-\pi(dx)\pi(dy) \| \leq 2(1+\pi(s))(1-\pi(s)) = 2(1-\{\pi(s)\}^2)$$

as claimed. □

Proof of Theorem 2. Let $P(x,dy)$ denote the Metropolis kernel, and let A_{n+1} be the event that the candidate for X_{n+1} is accepted. The conditional probability of A_{n+1}, given $X_n = x$ and $X_{n+1} = y$, is given by

$$P(A_{n+1}|X_n = x, X_{n+1} = y) = \frac{q(x,y)\alpha(x,y)\mu(dy)}{P(x,dy)}.$$

The conditional probability that $S_n = 1$, given $X_n = x$ and $X_{n+1} = y$, is therefore

$$\begin{aligned}P(S_n = 1|X_n = x, X_{n+1} = y) &= P(\{S_n = 1\} \cap A_{n+1}|X_n = x, X_{n+1} = y) \\ &= P\{S_n = 1|X_n = x, X_{n+1} = y, A_{n+1}\}P(A_{n+1}|X_n = x, X_{n+1} = y) \\ &= r_A(x,y)\frac{q(x,y)\alpha(x,y)\mu(dy)}{P(x,dy)} \\ &= \frac{s'(x)\nu'(dy)}{q(x,y)\alpha(x,y)\mu(dy)}\frac{q(x,y)\alpha(x,y)\mu(dy)}{P(x,dy)} \\ &= \frac{s'(x)\nu'(dy)}{P(x,dy)},\end{aligned}$$

which is the conditional regeneration probability (3.2) used in the construction of Theorem 1. □

Proof of Theorem 3. It is sufficient to show that

$$\min\left\{\frac{w(y)}{w(x)},1\right\} \geq \min\left\{\frac{c}{w(x)},1\right\}\min\left\{\frac{w(y)}{c},1\right\}$$

for all x and y and for any $c > 0$. To see this, note that if $c/w(x) \leq 1$, then

$$\begin{aligned}\min\left\{\frac{c}{w(x)},1\right\}\min\left\{\frac{w(y)}{c},1\right\} &= \frac{c}{w(x)}\min\left\{\frac{w(y)}{c},1\right\} \\ &= \min\left\{\frac{w(y)}{w(x)},\frac{c}{w(x)}\right\} \\ &\leq \min\left\{\frac{w(y)}{w(x)},1\right\},\end{aligned}$$

whereas if $c/w(x) \geq 1$, then

$$\min\left\{\frac{c}{w(x)},1\right\}\min\left\{\frac{w(y)}{c},1\right\} = \min\left\{\frac{w(y)}{c},1\right\} \leq \min\left\{\frac{w(y)}{w(x)},1\right\}$$

□

Proof of Proposition 3. Let $A \in \mathcal{E}$ and $x \in E$ be arbitrary, and set $B = A - \{x\}$. Since ν has no atoms,

$$\begin{aligned} s(x)\nu(A) &= s(x)\nu(B) \leq \int_B q(x,y)\alpha(x,y)\mu(dy) \\ &\leq \int_A q(x,y)\alpha(x,y)\mu(dy). \end{aligned}$$

This yields the required result. □

Proof of Proposition 4. Using the simplified conditional regeneration probability (4.5), the equilibrium regeneration probability is

$$\begin{aligned} \int\int \pi(x)p(x,y)r(x,y)\mu(dx)\mu(dy) &= \int\int \pi(x)f(y)\min\left\{\frac{w(y)}{c},1\right\}\min\left\{\frac{c}{w(x)},1\right\}\mu(dx)\mu(dy) \\ &= \int \min\{cf(x)\pi(x)\}\mu(dx)\int \min\left\{f(y),\frac{1}{c}\pi(y)\right\}\mu(dy) \\ &= c\left(\int \min\left\{f(x),\frac{1}{c}\pi(x)\right\}\mu(dx)\right)^2 \\ &= c\left(\int \min\left\{1,\frac{1}{c}\frac{\pi(x)}{f(x)}\right\}f(x)\mu(dx)\right)^2 \end{aligned}$$

□

References

ATHREYA, K. B. and NEY, P. (1978). A new approach to the limit theory of recurrent Markov chains. *Trans. Amer. Math. Soc.*, **245**, 493–501.

BESAG, J. and GREEN, P. J. (1993). Spatial statistics and Bayesian computation. *J. Roy. Statist. Soc. Ser. B*, **55**, to appear.

BOX, G. E. P. and TIAO, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Reading, MA: Addison-Wesley.

BRATLEY, P., FOX, B. L. and SCHRAGE, L. E. (1987). *A Guide to Simulation* (second ed.). New York, NY: Springer.

GAVER, D. P. and O'MUIRCHEARTAIGH, I. G. (1987). Robust empirical Bayes analysis of event rates. *Technometrics*, **29**, 1–15.

GELFAND, A. E., HILLS, S. E., RACINE-POON, A. and SMITH, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *J. Amer. Statist. Assoc.*, **85**, 972–985.

GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, **85**, 398–409.

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **??**, to appear.

GEWEKE, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, **57**, 1317–1339.

GEYER, C. J. (1992). On the convergence of Monte Carlo maximum likelihood calculations. Technical Report 571, School of Statistics, University of Minnesota.

GEYER, C. J. and THOMPSON, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. Roy. Statist. Soc. Ser. B*, **54**, 657–700.

GILKS, W. R., CLAYTON, D. G., SPIEGELHALTER, D. J., BEST, N. G., MCNEIL, A. J., SHARPLES, L. D. and KIRBY, A. J. (1993). Modeling complexity: Applications of Gibbs sampling in medicine. *J. Roy. Statist. Soc. Ser. B*, **55**, to appear.

HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

HILL, B. M. (1965). Inference about variance components in the one-way model. *J. Amer. Statist. Assoc.*, **60**, 806–825.

LIU, J., WONG, W. and KONG, A. (1991). Correlation structure and convergence rate of the Gibbs sampler with various scans. Technical Report 304, Department of Statistics, University of Chicago.

METROPOLIS, N., ROSENBLUTH, A., ROSENBLUTH, M., TELLER, A. and TELLER, E. (1953). Equations of state calculations by fast computing machines. *J. Chemical Physics*, **21**, 1087–1091.

NUMMELIN, E. (1978). A splitting technique for Harris recurrent markov chains. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, **43**, 309–318.

NUMMELIN, E. (1984). *General Irreducible Markov Chains and Non-Negative Operators.* Cambridge: Cambridge University Press.

POTTS, R. B. (1952). Some generalized order-disorder transformations. *Proc. Camb. Phil. Soc.*, **48**, 106–109.

RIPLEY, B. D. (1987). *Stochastic Simulation.* New York, NY: Wiley.

SMITH, A. F. M. and ROBERTS, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B*, **55**, to appear.

SWENDSEN, R. H. and WANG, J.-S. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.*, **58**, 86–88.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.*, **82**, 528–550.

TIERNEY, L. (1991a). Exploring posterior distributions using Markov chains. In *Computer Science and Statistics: 23rd Symposium on the Interface.* to appear.

TIERNEY, L. (1991b). Markov chains for exploring posterior distributions. Technical Report 560, School of Statistics, University of Minnesota.

YU, B. (1992). Density estimation in the L^∞ norm for depdendent data with applications to the Gibbs sampler. *Ann. Statist.*, **??**, to appear.